Primjena transformer neuronskih mreža za detekciju govora mržnje

Luka Utješinović

Univerzitet Crne Gore
Prirodno-matematički fakultet
Računarske nauke, master studije
Broj indeksa: 1/22

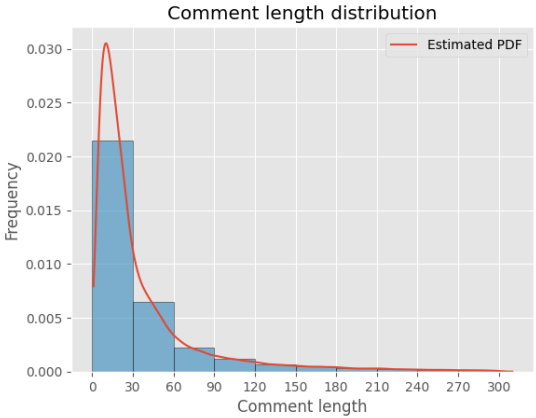
Dataset 1

- U postavci zadata 3 dataset-a:
  - **Hate Speech Detection Curated Dataset** dostupan na Kaggle platformi (link), rečenice razdvojene u dvije klase: hateful i non-hateful speech. Postoje dvije instance: prva sadrži $\approx$ 450 hiljada rečenica pri čemu non-hateful rečenice dominiraju, druga sadrži $\approx$ 700 hiljada rečenica, raspodjela klasa približno izjednačena.
  - **Toxic Comment Classification Challenge** dostupan na Kaggle patflromi (link), sadrži $\approx$ 150 hiljada rečenica razdvojenih u 6 kategorija: **toxic, severe_toxic, obscene, threat, insult, identity_hate**. Suštinski sadrži različite instance neprikladnog govora, u prvoj iteraciji ovaj dataset smo odbacili. Alternativa je da se sve prethodne rečenice označe kao hateful speech i pridruže prvom dataset-u.
  - **A Curated Hate Speech Dataset** dostupan na Mendeley platformi (link), gotovo identičan kao prvi dataset, tako da je i on odbačen.

Dataset 2

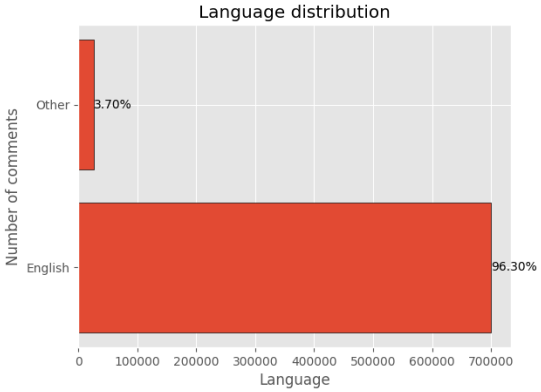
- Kako je nastala druga instanca prvog dataset-a:
  - Undersampling non-hateful rečenica
  - BERT augmentacija hateful rečenica - slučajno se odabere token iz rečenice, zamijeni se sa [MASK] tokenom i preda se pre-treniranom BERT modelu, iz dobijene vjerovatnosne raspodjele uzorkuju se nove rečenice.
  - Još jedan metod za augmentaciju korišćen - WordNet baza podataka.
- Dakle, rezultat je dataset od ≈ 700 hiljada rečenica, pri čemu su hateful i non-hateful klase približno jednako zastupljene.
- Preprocesiranje koje su vršili autori: uklanjanje ponovljenih whitespace-ova, lowercasing. Postoji argument da u ovom domenu lowercasing-om može da dođe do gubitka informacija!
- 80/20 train/validation split.

Dataset 3



- Svega petina rečenica sadrži broj riječi u opsegu (0, 30], ostale sadrže više od 30.
- Relativno duge rečenice - argument **protiv korišćenja rekurentnih neuronskih mreža** (vanishing/exploding gradient problem, opisano u seminarskom).

Dataset 4

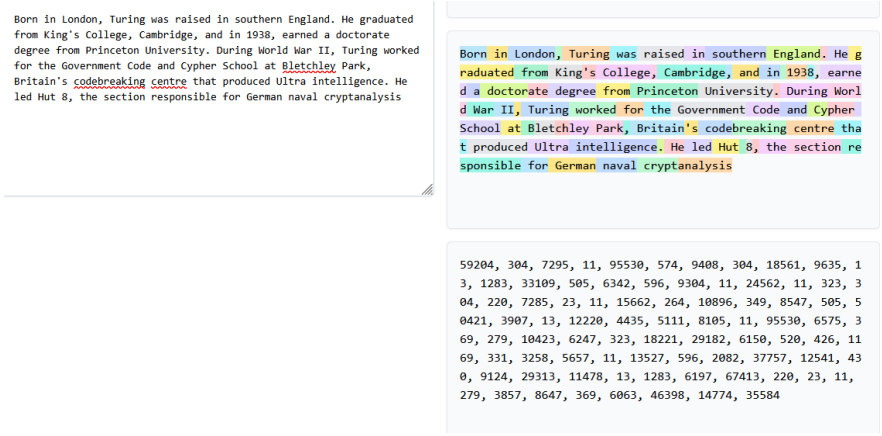


- Ručnom provjerom uočeno da su potencijalno zastupljeni različiti jezici. Za detekciju jezika rečenica korišćena je langid Python biblioteka.
- Da je raspodjela bila drugačija, bilo bi smisleno koristiti **multilingual model**. Ovi modeli detaljnije opisani u seminarskom!

Transformer neuronske mreže 1

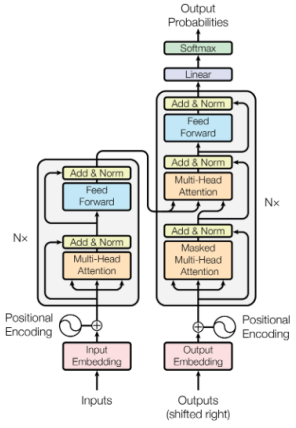
- Prije treniranja transformer modela treba napraviti vokabular od ulaznog dataset-a.
- Specijalni algoritmi za konstrukciju vokabulara na osnovu statističkih svojstava ulaznog dataset-a (**byte-pair encoding** opisan u seminarskom, **WordPiece, SentencePiece** ...)
- U praksi se često izbjegava bilo kakvo preprocesiranje teksta za transformer modele - ovaj dio "apsorbovan" od strane algoritma za tokenizaciju.
- Kompromis između veličine vokabulara i dužine tokenizovanih rečenica - u seminarskom detaljno opisan efekat algoritma za tokenizaciju!

Transformer neuronske mreže 2



Slika 1: Tokeni koje vraće BPE algortam za model GPT 3.5 Izvor: tiktokenizer

Transformer neuronske mreže 3



Slika 2: Izvorna arhitektura transformer neuronskih mreža. Popularizovana u članku [7]

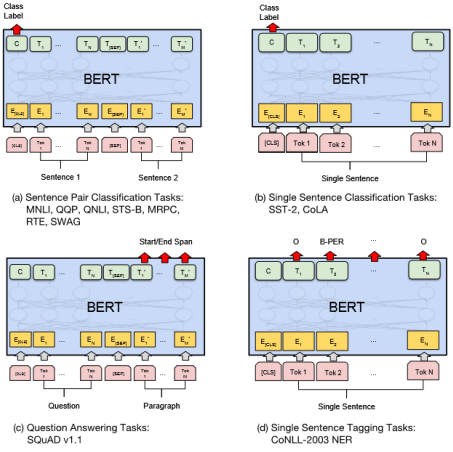
Transformer neuronske mreže 4

- Model koji se sastoji samo od enkoder komponente: klasifikacija teksta, **NER - Named Entity Recognition**, klasifikacija parova rečenica ...
- Model koji se sastoji samo od dekoder komponente: **ULM - Unconditional Language Models**: autoregresivno generisanje teksta. Sve popularne ChatBot aplikacije su dekoder transformer neuronske mreže.
- Modeli koji sadrže obije komponente: **CLM - Conditional Language Models** - izlaz dekoder komponente uslovljen izlazom enkoder komponente.
- **Sequence-to-sequence** paradigma: mašinsko prevođenje, sumarizacija teksta, ... Može da se primijeni i za klasifikaciju teksta!

BERT 1

- **BERT - Bidirectional Encoder Representation from Transformers**, [2]. Encoder-only transformer model.
- Pre-training vršen nad masivnim korpusom teksta koristeći **MLM - Masked Language Modelling** i **NSP - Next Sentence Prediction** metode. Detaljno opisano kako u seminarskom.
- Rezultujući model za ulazni token svake rečenice vraće **kontekstualnu reprezentaciju** (engl. **contextual embeddings**) - fine-tuning nad ovim reprezentacijama za problem od interesa.

BERT 2



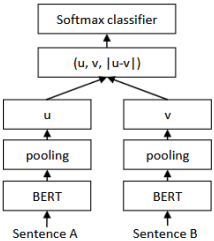
Slika 3: Ilustracija kako se vrši fine-tuning BERT modela za različite probleme. Pošto se radi samo o enkoder komponenti, nije moguće generativno modeliranje! Izvor: [2]

RoBERTa

- **RoBERTa - Robustly optimized BERT approach**, [4].
- Autori ovog članka uočili su određene bug-ove u originalnoj BERT implementaciji, koji su ispravljeni.
- Uklananje NSP iz pre-tranining procedure, povećanje **batch size parametra**, povećanje obima podataka za trening, ...
- Dovelo je do toga da rezultujući model daje bolje performanse. U našim eksperimentima vršićemo fine-tuning RoBERTa instance.

Sentence BERT 1

- Članak [6] pokrenuo je "lavinu" po pitanju primjene transformer enkoder modela u retrieval problemima.



Slika 4: Jedan metod učenja sematički bogatih reprezentacija kroz kontekstualne (Ro)BERT(a) reprezentacije. Izvor: [6]

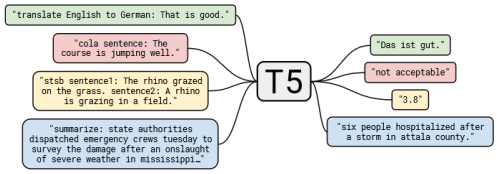
- Pooling operacija - suma, uzorački prosjek, maksimum po koordinatama, ...

Sentence BERT 2

- Rezultat je model koji proizvodi **semantički pogate reprezentacije** - semantički slične rečenice imaće slične reprezentacije, u odnosu na kosinusnu sličnost.
- Labela za hateful rečenicu: **Hatred, profanity, racial slurs.** Labela za non-hateful rečenicu: **Friendly, pleasantry, civility, gesture, levity.**
- Izračunati reprezentaciju ulazne rečenice i izračunati kosinusnu sličnost između reprezentacija prethodnih labela.
- Performanse ovog algoritma zavise od izbora labela za klase od interesa, biće opisan eksperiment kasnije.

T5

- **T5 - Text to Text Transfer Transformer**, [5]. Enkoder-dekoder transformer model koji svaki NLP problem tretira kao sequence-to-sequence problem.

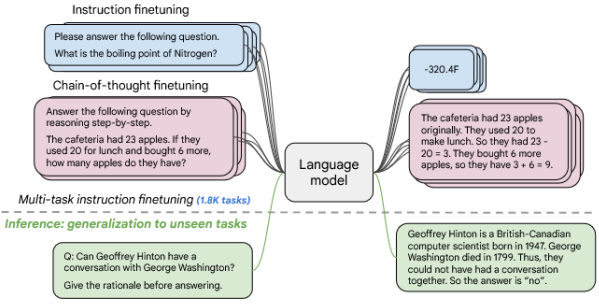


Slika 5: Ilustracija sequence-to-sequence paradigme.Izvor: [5]

- Ovo **nije** isto što i prompt-ovanje javno dostupnih LLM-ova, koji su decoder-only arhitekture!
- Za pre-training T5 modela koristi se **span corruption objective**, detaljno opisano u seminarskom kako funkcioniše.

FLAN T5

- Za naš dataset koristimo **FLAN T5**, [1].
- FLAN T5 $\approx$ pre-trained T5 instance uz sitne promjene u arhitekturi + fine-tuning nad dataset-ovima visokog kvaliteta



Slika 6: Ilustracija fine-tuning procedure za FLAN T5. Izvor, [1]

Metrika za evaluaciju 1

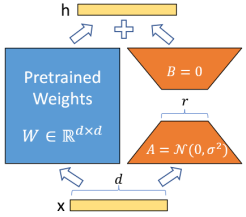
- RoBERTa - za zadatu rečenicu vraće vjerovatnosnu raspodjelu klasa.
- SentenceBERT - za zadatu rečenicu računamo reprezentaciju, i vraćemo klasu čija je labela najsličnija sa zadatom reprezentacijom. Moguće je da se pretvori u vjerovatnosnu raspodjelu - **softmax** operacija nad kosinusnim sličnostima.
- FLAN T5 - Enkoder dobija **Classify this sentence as hateful or non-hateful: rečenica**, dekoder obučavamo da na osnovu izlaza iz enkodera autoregresivno dekodira odgovarajuću labelu **Hateful** ili **Non-hateful**. Nije moguće pretvoriti u vjerovatnosnu raspodjelu koja bi bila uporediva sa prethodne dvije!

Metrika za evaluaciju 2

- Iako postoje argumenti da je **ROC-AUC** najrobusnija metrika za klasifikaciju (razmatra različite pragove vjerovatnoća), ovdje je neupotrebljiva zbog FLAN T5.
- Koristimo **macro-averaged** F_1 - težinski prosjek F_1 za sve klase pojedinačno, svaka klasa dobija težinu proporcionalnu sa njenom zastupljenošću.
- Pošto su u modifikovanom dataset-u klase približno jednako zastupljene, gotovo identično kao prosječni F_1

LoRA 1

- RoBERTa base broji $\approx 125M$ parametara, T5 base varijanta broji $\approx 220M$ parametara. Iako mogu da se treniraju na dostupnom hardveru, jedna epoha nad dataset-om od ≈ 700 hiljada rečenica bi traja predugo.
- Potreban kompromis - **LoRA - Low Rank Adaptation**



Slika 7: Ilustracija LoRA tehnike. Izvor: [3]

- Za svaku linearnu transformaciju,
  $\Delta W = W + \alpha AB, A, \in \mathbb{R}^{m \times r}, B \in \mathbb{R}^{r \times n}, r << max(m, n)$

Rezultati 1

- Prvo testiramo da li se Sentence (Ro)BERT(a) koji je retrieval model može koristiti za klasifikaciju, bez ikakvog treniranja. sentence-transformers Python biblioteka i all-distilroberta-v1 sentence transformer model. Uz prethodnu pomenutu biblioteku moguć fine-tuning ovog modela.

Labela za govor mržnje	Labela za govor koji nije govor mržnje	F_1
This is hateful speech.	This is not hateful speech	0.64
This is hateful and toxic content	**This is not hateful and non toxic content**	**0.71**
Hatred, profanity, racial slurs.	Friendly, pleasantry, civility, gesture, levity.	0.63

Tabela 1: Rezultati STS eksperimenta.

Rezultati 2

Model	Broj Param.	LoRA rang	LoRA broj param.	Broj epoha	Trening	F_1
all-distilroberta-v1	82M	–	–	–	–	0.71
google/flan-t5-base	248M	18	1.9M (130x manje)	4	43 h	0.86
roberta-base	125M	8	1.9M (65x manje)	8	26 h	0.9

Tabela 2: Upoređivanje retrieval modela sa RoBERTa i FLAN-T5 modelima.

- Hardver: **NVIDIA GTX 3060, 12GB VRAM + CuDA 11.8**
- **AdamW** algoritam za optimizaciju oba modela. Ostali hiperparametri opisani u seminarskom.

Tech Stack

- Vrlo vjerovatno rezultati mogu još da se poboljšaju: Bayesian hyperparameter optimization (Optuna Python biblioteka), transformer scaling laws, ...
- HuggingFace ekosistem:
  - transformers, pre-trenirani transformer modeli i njihov fine-tuning
  - sentence-transformers, SentenceBERT-like modeli
  - PEFT - Parameter Efficient Fine-Tuning implementira algoritme kao što su LoRA, Adapter injection, floating point quantization, ...
- Sve prethodne biblioteke imaju integraciju sa PyTorch bibliotekom, jedna od najpopularnijih biblioteka za **general-purpose deep learning**, korišćena i za ovaj projekat.

Bibliografija I

[1] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei.
Scaling instruction-finetuned language models, 2022.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova.
BERT: pre-training of deep bidirectional transformers for language understanding.
CoRR, abs/1810.04805, 2018.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen.
Lora: Low-rank adaptation of large language models.
CoRR, abs/2106.09685, 2021.

Bibliografija II

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov.
Roberta: A robustly optimized BERT pretraining approach.
CoRR, abs/1907.11692, 2019.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu.
Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[6] N. Reimers and I. Gurevych.
Sentence-bert: Sentence embeddings using siamese bert-networks.
CoRR, abs/1908.10084, 2019.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin.
Attention is all you need.
CoRR, abs/1706.03762, 2017.